A Kernel-Based Nonparametric Test for Anomaly Detection over Line Networks

Shaofeng Zou SZOU02@SYR.EDU
Department of EECS, Syracuse University, Syracuse, NY 13244

Yingbin Liang YLIANG06@SYR.EDU
Department of EECS, Syracuse University, Syracuse, NY 13244

H. Vincent Poor POOR@PRINCETON.EDU
Department of Electrical Engineering, Princeton University, Princeton, NJ 08544

Abstract

The nonparametric problem of detecting existence of an anomalous interval over a one-dimensional line network is studied. Nodes corresponding to an anomalous interval (if exists) receive samples generated by a distribution q, which is different from the distribution p that generates samples for other nodes. If anomalous interval does not exist, then all nodes receive samples generated by p. It is assumed that the distributions p and q are arbitrary, and are unknown. In order to detect whether an anomalous interval exists, a test is built based on mean embeddings of distributions into a reproducing kernel Hilbert space (RKHS) and the metric of maximum mean discrepancy (MMD). It is shown that as the network size n goes to infinity, if the minimum length of candidate anomalous intervals is larger than a threshold which has the order $O(\log n)$, the proposed test is asymptotically successful, i.e., the probability of detection error approaches zero asymptotically. An efficient algorithm to perform the test with substantial computational complexity reduction is proposed, and is shown to be asymptotically successful if the condition on the minimum length of candidate anomalous interval is satisfied. Numerical results are provided, which are consistent with the theoretical results.

1. Introduction

In this paper, we are interested in a type of problems, the goal of which is to detect existence of an anomalous object over a network. Each node in the network is associated with a random variable. An anomalous object, if exists, corresponds to a cluster of nodes in the network that take samples generated by a distribution q. All other nodes in the network take samples generated by the distribution p that is different from q. If an anomalous interval does not exist, then all nodes receive samples generated by p. Detection of no anomalous object (i.e., the null hypothesis H_0) against the anomalous event (i.e., hypothesis H_1) is a compound hypothesis testing problem due to the fact that the anomalous object may correspond to one of a number of candidate clusters in the network.

Such a problem models a variety of applications. For example, in sensor networks, sensors are deployed over a large range of space. These sensors take measurements from the environment in order to determine whether or not there is intrusion of an anomalous object. Such intrusion typically activates only a few sensors that cover a certain geometric area. An alarm is then triggered if the network detects an occurrence of intrusion based on sensors' measurements. Other applications can arise in detecting an anomalous segment of DNA sequences, detecting virus infection of computer networks, and detecting anomalous spot in images.

Detecting existence of a geometric object in large networks has been extensively studied in the literature. A number of studies focused on networks with nodes embedded in a lattice such as one dimensional line and square. In (Arias-Castro et al., 2005), the network is assumed to be embedded in a d-dimensional cube, and geometric structures such as line segments, disks, rectangles and ellipsoids associated with nonzero-mean Gaussian random variables need to be detected out of other nodes associated

with zero-mean Gaussian noise variables. A multiscale approach was proposed and its optimality was analyzed. In (Walther, 2010), detection of spatial clusters under the Bernoulli model over a two-dimensional space was studied, and a new calibration of the scan statistic was proposed, which results in optimal inference for spatial clusters. In (Pacifico et al., 2004), the problem of identifying a cluster of nodes with nonzero-mean values from zero-mean noise variables over a random field was studied.

Further generalization of the problem has also been studied, when network nodes are associated with a graph structure, and existence of an anomalous cluster or an anomalous subgraph of nodes needs to be detected. In (Arias-Castro et al., 2008), an unknown path corresponding to nonzero-mean variables needs to be detected out of zero-mean variables in a network with nodes connected in a graph. In (Addario-Berry et al., 2010), for various combinatorial and geometric structures of anomalous objects, conditions were established under which testing is possible or hopeless with a small risk. In (Arias-Castro et al., 2011), the cluster of anomalous nodes can either take certain geometric shapes or be connected as subgraphs. Such structures associated with nonzero-mean Gaussian variables need to be detected out of zero-mean variables. More recently, in (Sharpnack et al., 2013a;b), network properties of anomalous structures such as small cut size were incorporated in order to assist successful detection.

It can be seen that the majority of previous studies on this topic have taken parametric or semiparametric models on probability distributions, i.e., random variables are generated by known distributions such as Gaussian or Bernoulli distributions, or the two distributions are known to have mean shift. However, parametric models may not always hold in real applications. In many cases, distributions can be arbitrary, and may not be Gaussian or Bernoulli. They may not differ in mean either. Furthermore, distributions may not be known in advance. Hence, it is desirable to develop nonparametric tests that are distribution free.

1.1. Contributions

In contrast to previous studies, in this paper, we study the nonparametric model for anomalous interval detection, in which distributions can be arbitrary and unknown a priori. We focus on the problem of detecting existence of an anomalous interval over a one-dimensional line network. Although this is a simple network, it already captures the essence of the problem, and the same approach can be extended to studying more general network models.

In order to deal with the nonparametric model, we apply mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) (Berlinet & Thomas-Agnan, 2004; Sriperumbudur et al., 2010) (also see (Scholkopf & Smola, 2002; Hofmann et al., 2008) for an introduction of RKHS). The idea is to map probability distributions into a RKHS associated with an appropriate kernel such that distinguishing between two probabilities can be carried out by evaluating the distance between the corresponding mean embeddings in the RKHS. This is valid because the mapping is shown to be injective for certain kernels (Fukumizu et al., 2008; Sriperumbudur et al., 2008; Fukumizu et al., 2009; Sriperumbudur et al., 2010) such as Gaussian and Laplace kernels. The main advantage of such an approach is that the mean embedding of a distribution can be easily estimated based on samples. This approach has been applied to solving the two sample problem in (Gretton et al., 2007; 2012), in which the quantity of *maximum mean discrepancy (MMD)* was used as the metric of the distance between mean embeddings of two distributions.

Since the distributions can be arbitrary, it is in general difficult to exploit properties of the distributions such as mean shift to detect existence of an anomalous interval. Furthermore, as the network size becomes large (i.e., the number n of nodes goes to infinity), in contrast to parametric models in which the mean shift can scale with n, here it is necessary that the length of anomalous intervals (i.e., the number of samples over the anomalous distribution) is large enough in order for accurately identifying such an interval (see Remark 1 in Section 2). This implies that the scale of the anomalous object should enlarge as the detection range becomes larger in order for successful detection. Such a behavior also sets a clear difference of our nonparametric problem from previous studies of parametric models. Our goal is to characterize how the minimum length of candidate anomalous intervals should scale as the number of nodes goes to infinity in order to successfully detect existence of an anomalous interval.

We summarize our main contributions as follows.

(1) We address the nonparametric model of detecting existence of an anomalous interval over a line network. We identify the length of the anomalous interval as essential characteristic which must enlarge with n to guarantee successful detection in asymptotically large networks (see Remark 1 in Section 2). Hence, requiring the length of the anomalous interval to scale with n is necessary, not an artificial assumption.

(2) We build a distribution-free test using MMD based on kernel embeddings of distributions into RKHS. We analyze the performance guarantee of the proposed test, and show that as the network size n goes to infinity, if the minimum length of candidate anomalous intervals scales at the order $O(\log n)$[1] or larger, the proposed test can successfully

[1]In this paper, $f(n) = O(g(n))$ denotes $f(n)/g(n)$ converges to a constant as $n \to \infty$.

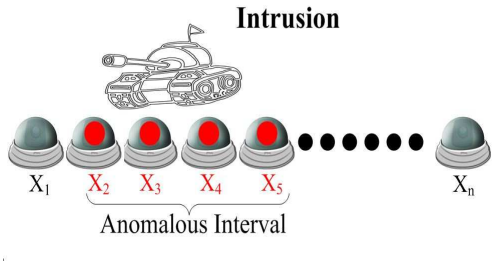

Figure 1. A line sensor network for intrusion detection

detect whether there exists an anomalous interval. Furthermore, we show that the test and the minimum length depends on prior knowledge of MMD of the two distributions.

(3) We adapt the multi-scale method in (Arias-Castro et al., 2005) and propose an efficient algorithm to perform the nonparametric test, which reduces the number of intervals for which MMD needs to be computed from the order $O(n^2)$ to $O(n^{1+\rho})$, for any $\rho > 0$. We further prove the performance guarantee for the proposed algorithm.

(4) We provide numerical results which are consistent with our theoretical assertions and demonstrate that the proposed test indeed provides guaranteed performance.

1.2. Organization

The rest of the paper is organized as follows. In Sections 2, we describe the problem formulation, define the performance measure, and clarify the difference of our problem from parametric models. In Section 3, we present our approach, algorithm, and main results for performance guarantee. In Section 4, we provide numerical results to demonstrate our theoretic results, and finally in Section 5, we conclude our paper with a few remarks on future work.

2. Problem Statement

We consider a line network, which consists of nodes $1, \ldots, n$, as shown in Figure 1. We use I to denote a subset of consecutive indices of nodes, which is referred to as an *interval*. Here, the length of an interval I refers to the cardinality of I, and is denoted by $|I|$. We assume that each node, say node i, is associated with a random variable, denoted by Y_i, for $i = 1, \ldots, n$. We use $\mathcal{I}_n$ to denote a set that contains all intervals over the network. We further denote the set of all candidate anomalous intervals as

$$\mathcal{I}_n^{(a)} = \{I \in \mathcal{I}_n : |I| \geq I_{\min}\} \tag{1}$$

where $I_{\min}$ denotes the minimum length of candidate anomalous intervals. The reason of imposing such a minimum length requirement is explained in Remark 1.

We consider two hypotheses about the distributions of the line network. For the *null hypothesis* H_0, Y_i for $i = 1, \ldots, n$ are identical and independently distributed (i.i.d.) random variables, and are generated from a distribution p. For the *alternative hypothesis* H_1, there exists an interval $I \in \mathcal{I}_n^{(a)}$ over which Y_i are i.i.d. and are generated from a distribution $q \neq p$ for all $i \in I$, and otherwise, Y_i are i.i.d. and generated from the distribution p. We further assume that under both hypotheses , each node generates only one sample. Putting the problem into a context, H_0 models the scenario when the observations Y_i are background noise, and H_1 models the scenario when some Y_i (for $i \in I$) are observations activated by an anomalous object.

In contrast to previous work, we assume that the distributions p and q are arbitrary and are unknown a priori. Instead, one sample X_i is independently generated from the distribution p for each node as a reference sample for the null hypothesis. This is reasonable because in practical scenarios, systems typically start under H_0 and it is not difficult to collect samples at this stage. For example, in the case of intrusion detection, the system is typically set up and activated before any intrusion occurs. Hence, samples collected at such an initial state can serve as reference of the null hypothesis.

For this problem, we are interested in the asymptotical scenario, in which the number of nodes goes to infinity, i.e., $n \to \infty$. The performance of a test for such a system is captured by the two types of errors. The *type I error* refers to the event that samples are generated from the null hypothesis, but the detector determines an anomalous event occurs. We denote the probability of such an event as $P(H_1|H_0)$, or $P_{H_0}(error)$. The *type II error* refers to the case that an anomalous event occurs but the detector claims that the sample are generated from the null hypothesis. We denote the probability of such an event as $P(H_0|H_1)$, or $P_{H_1}(error)$.

Definition 1. *A test is said to be asymptotically successful if*

$$\lim_{n\to\infty} P(H_1|H_0) + P(H_0|H_1) \to 0. \tag{2}$$

This hypothesis testing problem has a compound nature in that H_1 includes events corresponding to all candidate intervals where an anomalous object can locate, i.e., for all $I \in \mathcal{I}_n^{(a)}$. In general, an anomalous interval with smaller length is more difficult to detect due to the small number of samples from the anomalous distribution q. As $n \to \infty$, the total number of intervals goes to infinity in the order of $O(n^2)$. In this case, in order for successful detection, each candidate anomalous interval should provide more accurate information about the corresponding distribution. This requires that the length of candidate anomalous intervals enlarge with n. This suggests that as the network becomes larger, it can detect only a large enough anomalous object.

Remark 1. *We argue that it is necessary for the minimum length $I_{\min}$ of candidate anomalous intervals to enlarge to infinity as $n \to \infty$ in order to guarantee asymptotically successful detection in nonparametric model. This is because in the nonparametric model, the distributions p and q are fixed as n changes. Now suppose p and q are both Gaussian but with different mean values. Since mean values do not scale with n in our model, following (Arias-Castro et al., 2005) Theorem 2.3, no test can be asymptotically successful if $I_{\min}$ is bounded. Therefore, no distribution-free test exists if the minimum length $I_{\min}$ is bounded as $n \to \infty$.*

Therefore, our goal in this problem is to characterize how the minimum length $I_{\min}$ of candidate anomalous intervals should scale with the network size (i.e., the number of nodes) in order for a detector to successfully distinguish between the two hypotheses.

2.1. Comparison with Parametric Models

In this subsection, we compare our nonparametric model with the parametric model. We take the Gaussian model studied in (Arias-Castro et al., 2005) as an example, in which the anomalous distribution q differs from p in mean. It is required that the mean difference of the two distributions enlarge with the network size n in order for differentiating the two hypotheses. Otherwise, as the network size approaches infinity (and correspondingly, the number of intervals becomes larger), it is likely that one interval happens to have samples with large values even under the null hypothesis, which is likely to cause detection error.

For the nonparametric model studied in this paper, since the distributions are unknown and can be arbitrary, there is no particular parameter such as the mean that captures distinction between the two distributions. In fact, the means of the two distributions can be the same. Therefore, in this case, in order to distinguish between the two hypotheses in large networks, as argued in Remark 1, it is necessary that the samples of the anomalous distribution are large enough in order to well identify such a distribution. Furthermore, such a setting also complements the parametric model in the sense that if the mean of the Gaussian distribution does not scale with the network size, as long as the anomalous object is large enough, successful detection is still possible.

3. Main Results

In this section, we first introduce the approach that we use based on MMD of kernel embeddings of distributions. We then present construction of a test for our problem and theoretical analysis of this test. Finally, we present an algorithm conducting the test with low computational complexity.

3.1. Introduction of MMD

We provide a brief introduction of the idea on mean embedding of distributions into RKHS (Berlinet & Thomas-Agnan, 2004; Sriperumbudur et al., 2010) and the metric of MMD. Suppose $\mathcal{P}$ includes a class of probability distributions, and suppose $\mathcal{H}$ is the RKHS with an associated kernel $k(\cdot,\cdot)$. We define a mapping from $\mathcal{P}$ to $\mathcal{H}$ such that each distribution $p \in \mathcal{P}$ is mapped into an element in $\mathcal{H}$ as follows

$$\mu_p(\cdot) = \mathbb{E}_p[k(\cdot, x)] = \int k(\cdot, x) dp(x).$$

Here, $\mu_p(\cdot)$ is referred to as the *mean embedding* of the distribution p into the Hilbert space $\mathcal{H}$. Due to the reproducing property of $\mathcal{H}$, it is clear that $\mathbb{E}_p[f] = \langle \mu_p, f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

It is desirable that the embedding is *injective* such that each $p \in \mathcal{P}$ is mapped to a unique element $\mu_p \in \mathcal{H}$. It has been shown in (Fukumizu et al., 2008; Sriperumbudur et al., 2008; Fukumizu et al., 2009; Sriperumbudur et al., 2010) that for many RKHSs such as those associated with Gaussian and Laplace kernels, the mean embedding is injective. In this way, many machine learning problems with unknown distributions can be solved by studying mean embeddings of probability distributions without actually estimating the distributions, e.g., (Song et al., 2013)(Song et al., 2011b;a; Smola et al., 2007). For example, two-sample problem can be solved by comparing the mean embeddings of two distributions as in (Gretton et al., 2012). In order to distinguish between two distributions p and q, (Gretton et al., 2012) introduced the following quantity of maximum mean discrepancy (MMD) based on the mean embeddings μ_p and μ_q of p and q in RKHS:

$$\text{MMD}[p, q] := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{3}$$

It is also shown that

$$\text{MMD}[p, q] = \sup_{\|f\|_{\mathcal{H}} \le 1} \mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)].$$

Namely, $\text{MMD}[p, q]$ achieves the maximum of the mean difference of a function between the two distributions over all unit-norm functions in the RKHS $\mathcal{H}$.

Due to the reproducing property of kernel, it can be easily shown that

$$\begin{aligned}\text{MMD}^2[p, q] =& \mathbb{E}_{x,x'}[k(x, x')] - 2\mathbb{E}_{x,y}[k(x, y)] \\ &+ \mathbb{E}_{y,y'}[k(y, y')],\end{aligned} \tag{4}$$

where x and x' have independent but the same distribution p, and y and y' have independent but the same distribution q. An unbiased estimate of $\text{MMD}^2[p, q]$ based on n sam-

ples of x and m samples of y is given by

$$\mathrm{MMD}_u^2[X,Y] = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(y_i,y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i,y_j). \tag{5}$$

We note that other estimators of the $\mathrm{MMD}^2[p,q]$ are also available, which can be used for our problem. In this paper, we focus on the unbiased estimate given above to convey the central idea.

3.2. Test and Performance Analysis

We construct our test using the unbiased estimate $\mathrm{MMD}_u^2[X,Y]$ of $\mathrm{MMD}^2[p,q]$ given in (5). In particular, for each interval $I \in \mathcal{I}_n^{(a)}$, we compute $\mathrm{MMD}_{u,I}^2[X,Y]$ based on samples $(y_j,\ j \in I)$ and the reference sequence $(x_1,\ldots,x_n)$ generated by p as follows:

$$\mathrm{MMD}_{u,I}^2[X,Y] = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j) + \frac{1}{|I|(|I|-1)}\sum_{i\in I}\sum_{j\neq i, j\in I} k(y_i,y_j) - \frac{2}{n|I|}\sum_{i=1}^{n}\sum_{j\in I} k(x_i,y_j). \tag{6}$$

If there exists an anomalous interval I, we expect the corresponding $\mathrm{MMD}_{u,I}^2[X,Y]$ to be large, because the sequence of the anomalous interval is generated by a distribution q differently from the reference sequence. Otherwise, under null hypothesis $\mathrm{MMD}_{u,I}^2[X,Y]$ should be small for all candidate I. Hence, we build our test as follows.

$$\max_{I: I\in\mathcal{I}_n^{(a)}} \mathrm{MMD}_{u,I}^2(X,Y) \begin{cases} \geq t, & \text{determine } H_1 \\ < t, & \text{determine } H_0 \end{cases} \tag{7}$$

where t is a threshold parameter, which is determined in Corollaries 1 and 2.

We next analyze the performance of the above test. The following theorem characterizes how the minimum length $I_{\min}$ of candidate anomalous intervals scales with the network size n so that the test (7) is guaranteed to be asymptotically successful.

Theorem 1. *Suppose the test (7) is applied to the nonparametric problem given in Section 2. Further assume that the kernel in the test satisfies $0 \leq k(x,y) \leq K$ for all (x,y). Then the test (7) is asymptotically successful if*

$$I_{\min} \geq \frac{16K^2(1+\eta)}{t^2}\log n \tag{8}$$

where η is any positive constant, and t is the threshold of the test that satisfies $t < \mathrm{MMD}^2[p,q]$.

We note that the boundedness assumption on $k(x,y)$ is satisfied for many kernels such as Gaussian and Laplace kernels. We further note that the above theorem implies that the minimum length $I_{\min}$ can be in the order $O(\log n)$. Hence, the number of candidate anomalous intervals in the set $\mathcal{I}_n^{(a)}$ is in the order $O(n^2)$, which is the same as the number of all intervals. Hence, in the order sense, not many intervals are excluded from being anomalous.

Theorem 1 requires that the threshold t in the test (7) to be less than $\mathrm{MMD}^2[p,q]$. The information of $\mathrm{MMD}^2[p,q]$ may or may not be available depending on specific applications. In some cases, samples from anomalous events are also collected, and hence $\mathrm{MMD}^2[p,q]$ can be estimated reasonably well by (5). In such cases, the threshold t can be set as a constant smaller than $\mathrm{MMD}^2[p,q]$. On the other hand, if samples from q are not available, then the threshold t needs to scale to zero as n gets large in order to be asymptotically smaller than $\mathrm{MMD}^2[p,q]$. We summarize these two cases in the following corollaries.

Corollary 1. *If $\mathrm{MMD}^2[p,q]$ is known a priori, then set the threshold t as $t = (1-\delta)\mathrm{MMD}^2[p,q]$ for any $0 < \delta < 1$. In this case, the test (7) is asymptotically successful if*

$$I_{\min} \geq \frac{16K^2(1+\eta')}{\mathrm{MMD}^4[p,q]}\log n \tag{9}$$

where η' is any positive constant.

Corollary 1 follows directly from Theorem 1 by setting $\eta' = \frac{1+\eta}{(1-\delta)^2} - 1$.

Corollary 2. *If $\mathrm{MMD}^2[p,q]$ is unknown a priori, then set the threshold t to scale with n such that $\lim_{n\to\infty} t_n = 0$. In this case, the test (7) is asymptotically successful if*

$$I_{\min} \geq \frac{16K^2(1+\eta)}{t_n^2}\log n. \tag{10}$$

Corollary 2 follows directly from Theorem 1 by noting that $t_n < \mathrm{MMD}^2[p,q]$ for large n.

We note that Corollary 2 holds for any t_n that satisfies $\lim_{n\to\infty} t_n = 0$. It is clear from Corollary 2 that for the case when $\mathrm{MMD}^2[p,q]$ is unknown, the minimum length $I_{\min}$ is strictly larger than the order $O(\log n)$.

We further note that the above two corollaries demonstrate that the prior knowledge about $\mathrm{MMD}^2[p,q]$ is very important for network capability in identifying anomalous events. If $\mathrm{MMD}^2[p,q]$ is known, then the network can resolve an anomalous object with the length in the order $O(\log n)$. However, if such knowledge is unknown, the network can resolve only bigger anomalous objects with the length larger than $O(\log n)$.

3.3. Outline of Proof of Theorem 1

We first introduce the definition of dyadic intervals and their properties (Arias-Castro et al., 2005), which are useful for the proof of Theorem 1 and for understanding Algorithm 1 in Section 3.4. For convenience, assume that $n = 2^J$, where J is an integer, and define the *dyadic intervals* as

$$I_{j,k} = \{k2^j, \ldots, (k+1)2^j - 1\} \quad \text{for } 0 \le j \le \log_2(n), \text{ and } 0 \le k \le \frac{n}{2^j}. \tag{11}$$

We let $\mathcal{I}_n^{(d)}$ denote the set that consists of all dyadic intervals. It can be shown in (Arias-Castro et al., 2005) that for any interval I,

$$\frac{|I^{(d)}|}{|I|} \ge \frac{1}{4}$$

where $I^{(d)}$ is the maximum dyadic interval contained in I.

We next define the *l-level extensions of a dyadic interval* $I_{j,k}$ as follows. Starting from the interval $I_{j,k}$ or the union of $I_{j,k}$ and $I_{j,k+1}$ where k is odd, at level $q = 1, \ldots, l$, attach dyadic intervals of length $2^{-q}|I_{j,k}|$ at either, or both ends of interval resulted from the previous step, or do nothing. Let $\mathcal{J}_{n,l}(I)$ to denote the set that includes all l-level extensions of a dyadic interval I. Let $\mathcal{J}_{n,l} = \cup_{I \in \mathcal{I}_n^{(d)}} \mathcal{J}_{n,l}(I)$. We note the following useful properties.

Lemma 1. *(Arias-Castro et al., 2005)*

(1) $|\mathcal{J}_{n,l}| \le n4^{l+1}$;

(2) For any interval I and the corresponding maximum dyadic interval $I^{(d)}$ contained in I, there exists one interval J in the l-level extension $\mathcal{J}_{n,l}(I^{(d)})$ of $I^{(d)}$ such that $|I| - |J| \le 2^{-(l-1)}|I^{(d)}|$.

Outline of Proof of Theorem 1. We now provide the main idea to prove Theorem 1 with the complete proof provided in supplementary materials.

For each set $I \in \mathcal{I}_n^{(a)}$, we find the corresponding set $J_I \in \mathcal{J}_{n,l}(I^{(d)})$ that satisfies Lemma 1 (2). We then collect all such intervals J_I of l-level extensions into a set $\mathcal{J}_{n,l}^{(a)}$. The idea of the proof is to use $\max_{J \in \mathcal{J}_{n,l}^{(a)}} \text{MMD}^2_{u,J}[X, Y]$ as a good approximation of the true test in (7). Since the number of intervals in $\mathcal{J}_{n,l}^{(a)}$ is much smaller than the number of intervals in $\mathcal{I}_n^{(a)}$, a test based on this approximation helps to tighten the result. Based on this idea, under H_0, we bound $P_{H_0}(error)$ as,

$$P_{H_0}(error) = \exp\left(2\log n - \frac{\epsilon^2 2^l I_{\min}}{c_1 + c_2 + c_3}\right) + \exp\left(\log n + (l+1)\log 4 - \frac{(t-\epsilon)^2(1-2^{1-l})I_{\min}}{16K^2}\right) \tag{12}$$

where c_1, c_2, c_3 are constants. We further set

$$\epsilon = (1-\beta)t$$

where $0 < \beta < 1$ is a constant. It can be shown that there exist β close enough to 1 and l large enough (but a constant) such that (12) implies that

$$P_{H_0}(error) \to \infty$$

as $n \to \infty$ if $I_{\min} > \frac{16K^2(1+\eta)}{t^2}\log n$.

Under H_1, suppose $\hat{I}$ is the anomalous interval. Using the fact that $t < \text{MMD}^2[p, q]$, we have the following bound

$$P_{H_1}(error) \le \exp\left(-\frac{(\text{MMD}^2[p,q] - t)^2 I_{\min}}{16K^2}\right) \tag{13}$$

which converges to zero as $n \to \infty$, if $I_{\min} > \frac{16K^2(1+\eta)}{t^2}\log n$.

□

3.4. An Efficient Algorithm

In this subsection, we describe an efficient algorithm to perform the proposed test (7). In general, since the number of all intervals with length larger than $I_{\min}$ has an order $O(n^2)$, the test (7) requires to compute $\text{MMD}^2_{u,I}[X, Y]$ for $O(n^2)$ intervals. We next provide Algorithm 1 that computes $\text{MMD}^2_{u,I}[X, Y]$ for only $O(n^{1+\rho})$ intervals for any $\rho > 0$. This algorithm adapts the multi-scale method in (Arias-Castro et al., 2005) for parametric models.

The basic idea of the algorithm is to use the set of dyadic intervals and their extensions (as introduced in Section 3.3) such that $\text{MMD}^2_{u,I}[X, Y]$ over any interval is well approximated by an interval in such a set. A key property of such a set is that its cardinality is in the order $O(n^{1+\rho})$ for any $\rho > 0$, which reduces computation of $\text{MMD}^2_{u,I}[X, Y]$ for only $O(n^{1+\rho})$ intervals.

The following theorem provides the performance guarantee for Algorithm 1.

Theorem 2. *Algorithm 1 is asymptotically successful with computation of $\text{MMD}^2_{u,I}[X, Y]$ for the order $O(n^{1+\rho})$ intervals for any $\rho > 0$.*

Algorithm 1 Detect Existence of an Anomalous Interval

Input: n; $t = t_n \to 0$; $t' < t$; $\eta > 0$; $I_{\min} \geq \frac{16K^2(1+\eta)}{t^2} \log n$, $\delta > \frac{\eta}{2}$; and $l = \lceil \log_2 \left(\frac{1+\eta}{\eta} \right) + 2 \rceil$.

Output:

- Construct set $\mathcal{I}_n^{(p)} = \{I \in \mathcal{I}_n^{(d)} : |I| \geq \frac{I_{\min}}{4}$, and $\text{MMD}^2_{u,I}[X,Y] \geq t'\}$;
- If $|\mathcal{I}_n^{(p)}| > n^{1-\frac{t'^2}{4t^2}(1+\eta)+\delta}$, then determine H_1;
- If $\max_{I \in \mathcal{I}_n^{(p)}} \text{MMD}^2_{u,I}[X,Y] > \frac{2t}{\sqrt{1+\eta/2}}$, then determine H_1;
- Construct set $\mathcal{J}_{n,l}^{(p)}$ that includes level-l extension of all intervals in $\mathcal{I}_n^{(p)}$ with length larger than $\frac{16K^2(1+\eta/2)}{t^2} \log n$;
- If $\max_{I \in \mathcal{J}_{n,l}^{(p)}} \text{MMD}^2_{u,I}[X,Y] > t$, then determine H_1;
- Otherwise, determine H_0.

Outline of Proof of Theorem 2. We describe the main idea of the proof here with the complete proof provided in supplementary materials. We show that Algorithm 1 guarantees asymptotically small probability of detection error as $n \to \infty$.

Under H_1, due to the algorithm, the error event is contained in the event $\max_{I \in \mathcal{J}_{n,l}^{(p)}} \text{MMD}^2_{u,I}[X,Y] < t$. Based on this fact, we can show that $P_{H_1}(error)$ converges to zero as n goes to infinity.

Under H_0, there are three cases of determining H_1 in the algorithm which cause errors: (1) Case 1 error occurs if $|\mathcal{I}_n^{(p)}| > n^{1-\frac{t'^2}{4t^2}(1+\eta)+\delta}$; (2) Case 2 error occurs if $\max_{I \in \mathcal{I}_n^{(p)}} \text{MMD}^2_{u,I}[X,Y] > \frac{2t}{\sqrt{1+\eta/2}}$; and (3) Case 3 error occurs if $\max_{I \in \mathcal{J}_{n,l}^{(p)}} \text{MMD}^2_{u,I}[X,Y] > t$. For all three cases, we show that the probability of error converges to zero as n goes to infinity. □

4. Numerical Results

In this section, we provide five numerical tests to demonstrate our theoretical results for both cases with known and unknown MMD$[p,q]$. In these numerical results, we average the two types of errors as $P_e = \frac{1}{2}\left(P_{H_1}(\text{error}) + P_{H_0}(\text{error})\right)$.

The first three tests study how $I_{\min}$ scales with the network size n in order to guarantee asymptotically small probability of error for three scenarios. For these tests, we assume that a good estimate of MMD$[p,q]$ is available.

In test 1, the distribution p is $\mathcal{N}(0, 0.5)$, and the anomalous distribution q is a mixture of two Gaussian distributions $\mathcal{N}(-2, 0.5)$ and $\mathcal{N}(2, 0.5)$ with equal probability. This test models the case that an anomalous object activates Bernoulli distribution. For detection, we use Gaussian kernel $k(x,x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$ with $\sigma = 1$, and we set the threshold $t = 0.25$. We run the test for networks with sizes $n = 40, 100, 200, 300, 500$, respectively.

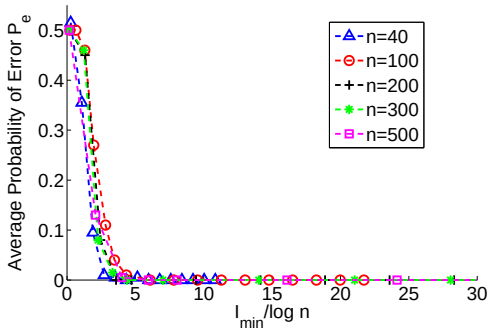

Figure 2. Test 1: Performance of detecting an anomalous interval distributed as a mixture of two Gaussian distributions out of a Gaussian distributed line network with Gaussian kernel and known MMD

In Figure 2, we plot how the average probability of error changes with the minimum length $I_{\min}$. For illustration convenience, we normalize $I_{\min}$ by $\log n$. It can be seen that when $I_{\min}/\log n$ is above a certain threshold, the probability of error converges to zero, which is consistent with our theoretical results. Furthermore, for different n, all curves drop to zero almost at the same threshold. Such behavior also agrees with Theorem 1, which states that the threshold depends only on the bound of the kernel and the threshold of the test, and these parameters are the same for all curves.

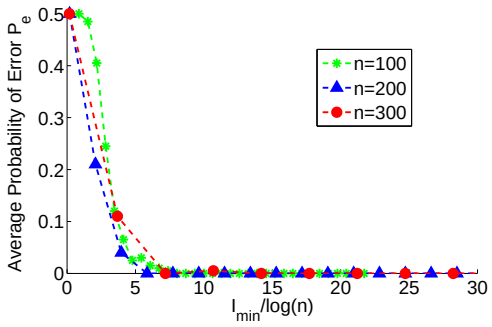

Figure 3. Test 2: Performance of detecting an anomalous interval with Gaussian distribution that has a different variance from other nodes in a line network with Gaussian kernel and known MMD

In test 2, distributions p and q are respectively chosen to be $\mathcal{N}(0,1)$ and $\mathcal{N}(0,4)$, i.e., they have the same mean but different variances. We use Gaussian kernel with $\sigma = 1$

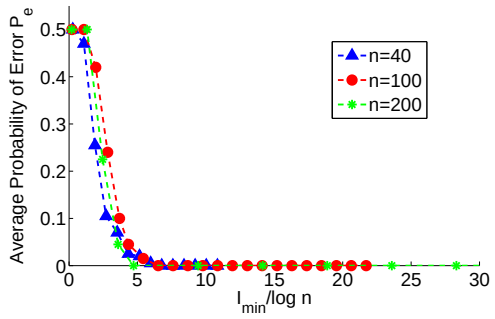

Figure 4. Test 3: Performance of detecting an anomalous interval distributed as a mixture of two Gaussian distributions out of a Gaussian distributed line network with Laplace kernel and known MMD

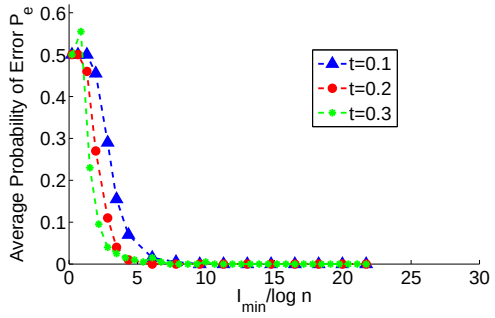

Figure 5. Test 4: Impact of threshold t in the test on $I_{\min}$ to guarantee asymptotically small probability of error

for the test, and set threshold $t = 0.1$. We run the test for networks with sizes $n = 100, 200, 300$. Figure 3 plots how the average probability of error changes with the minimum length $I_{\min}$, and demonstrates a behavior similar to test 1.

In test 3, the distributions p and q are the same as test 1. Instead of using Gaussian kernel for the test, we use Laplace kernel $k(x, x') = \exp\left(-\frac{|x-x'|}{2\sigma}\right)$ with $\sigma = 1$. We run the test for networks with sizes $n = 40, 100, 200$. In Figure 4, we plot the average probability of error versus $\frac{I_{\min}}{\log n}$. It can be seen that although this test applies a different kernel, the performance is similar to Figures 2 and 3 for tests 1 and 2, respectively.

Summarizing the results from tests 1, 2 and 3, it is clear that our approach is robust to changes in distributions and changes in kernels. In particular, all tests demonstrate the threshold scaling behavior of the minimum length $I_{\min}$ to guarantee asymptotically small probability of error.

In test 4, we study how the threshold t affects the scaling behavior of $I_{\min}$ with the network size. We choose the same distributions p and q as test 1. We also use Gaussian kernel with $\sigma = 1$ for our test. We study a network with

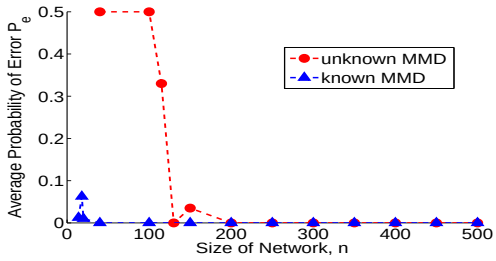

Figure 6. Test 5: Comparison of performance between tests with unknown MMD and known MMD

size $n = 100$. We run the test for $t = 0.1, 0.2, 0.3$. In Figure 5, we plot the average probability of error versus $\frac{I_{\min}}{\log n}$ corresponding to different values of t. Although the curves exhibit the behavior similar to that in the first three tests, the probabilities of error do not drop at the same threshold on $\frac{I_{\min}}{\log n}$. It can be seen that as t enlarges, the dropping threshold on $\frac{I_{\min}}{\log n}$ gets smaller, implying that the network can detect smaller anomalous object. This is consistent with Theorem 1, which suggests that the threshold on $I_{\min}$ to guarantee asymptotically small probability of error is inversely proportional to t^2.

In test 5, we study the case when the MMD is unknown, and compare its performance with the case when the MMD is known. We choose the distribution p to be $\mathcal{N}(0, 0.5)$, and choose the distribution q to be a mixture of two Laplace distributions with the same variance 0.5 and different means -3 and 3 equally likely. We use Gaussian kernel with parameter $\sigma = 0.9$. Since the MMD is unknown, we set the threshold t to change with n as $t_n = 4\sqrt{\frac{\log n}{n^{0.9}}}$, which goes to zero as n goes to infinity. As the network size n changes, we set the minimum length of anomalous intervals as $I_{\min} = \lceil n^{0.9} \rceil$ suggested by Theorem (1). We also run a comparison test with MMD known in advance, for which we set the threshold $t = 0.1$. For a fair comparison, we also set $I_{\min} = \lceil n^{0.9} \rceil$.

In Figure 6, we plot how the average probability of error changes as a function of the network size n with unknown MMD. It can be seen that as n becomes large, the probability of error goes to zero demonstrating that our test is asymptotically successful. This also agrees with Theorem 1 because we have chosen the minimum length to satisfy (8). It can also be seen from Figure 6 that the probability of error for the case with known MMD converges much faster than the case with unknown MMD, demonstrating the importance of prior knowledge of MMD.

5. Conclusions

In this paper, we investigated the nonparametric problem of detecting existence of an anomalous interval over a line

network, in which both normal and anomalous distributions are arbitrary and unknown a priori. We built a distribution free test using the MMD to measure the distance between the mean embeddings of two distributions into a RKHS. We showed that if the minimum length of candidate anomalous intervals is above a certain threshold in the order larger than or equal to $O(\log n)$, then our test is asymptotically successful (i.e., the probability of error converges to zero) as the network size n goes to infinity. Furthermore, we proposed an efficient algorithm to perform the test with reduced complexity in computing MMD, and showed that the algorithm is guaranteed to be asymptotically successful. Our results demonstrate that the metric of the MMD is a powerful technique for studying nonparametric problems and for building distribution free test. We believe such an approach can be applied to various detection problems involving distinguishing among distributions. It is also of interest to study tests based on other estimators of the MMD, and compare the performances of these tests.

References

Addario-Berry, L., Broutin, N., Devroye, L., and Lugosi, G. On combinatorial testing problems. *Ann. Statist.*, 38 (5):3063–3092, 2010.

Arias-Castro, E., Donoho, D. L., and Huo, X. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory*, 51(7):2402–2425, July 2005.

Arias-Castro, E., Candes, E. J., Helgason, H., and Zeitouni, O. Searching for a trail of evidence in a maze. *Ann. Statist.*, 36(4):1726–1757, 2008.

Arias-Castro, E., Candes, E. J., and Durand, A. Detection of an anomalous cluster in a network. *Ann. Statist.*, 39 (1):278–304, 2011.

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2008.

Fukumizu, K., Sriperumbudur, B., Gretton, A., and Schölkopf, B. Characteristic kernels on groups and semigroups. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2009.

Gretton, A., Borgwardt, K., Rasch, M., Schökopf, B., and Smola, A. A kernel method for the two-sample-problem. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2007.

Gretton, A., Borgwardt, K., Rasch, M., Schökopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 2008.

Pacifico, P. M., Genovese, C., Verdinelli, I., and Wasserman, L. False discovery control for random fields. *J. Amer. Stat. Assoc.*, 99:1002–1014, 2004.

Scholkopf, B. and Smola, A. *Learning with Kernels*. MA: MIT press, 2002.

Sharpnack, J., Rinaldo, A., and Singh, A. Changepoint detection over graphs with the spectral scan statistic. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, Scottsdale, AZ, May 2013a.

Sharpnack, J., Rinaldo, A., and Singh, A. Detecting activations over graphs using spanning tree wavelet bases. In *Artificial Intelligence and Statistics (AISTATS)*, Scottsdale, AZ, May 2013b.

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. Algorithmic Learning Theory, E. Takimoto(Eds.), *Lecture Notes on Computer Science*, Springer, 2007.

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. Kernel belief propagation. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, 2011a.

Song, L., Parikh, A., and Xing, E. P. Kernel embeddings of latent tree graphical models. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2011b.

Song, L., Gretton, A., and Fukumizu, K. Kernel embeddings of conditional distributions. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Scholkopf, B. Injective Hilbert space embeddings of probability measures. In *Proc. Annual Conference on Learning Theory (COLT)*, 2008.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Scholkopf, B. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

Walther, G. Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.*, 38(2):1010–1033, 2010.